Analysis of the Banff challenge 2a problems using the method of fractional event counting

Stefan Schmitt, DESY

December 10, 2010

1 The method of fractional event counting

The Banff 2a discovery challenge problems are analysed using the method of fractional event counting [1]. Weights are assigned for a mark x, where the weights are given by

$$w(x) = \frac{S_0(x)}{D(x)} - k_1 \sum_m A_m \frac{\sigma_m^{SB}(x)}{D(x)} - k_2 \sum_m B_m \frac{\sigma_m^B(x)}{D(x)}, \text{ where} \quad (1)$$

$$D(x) = k_1(S_0(x) + B_0(x) + \sigma_{unc}^S(x)^2 + \sigma_{unc}^B(x)^2) + k_2(B_0(x) + \sigma_{unc}^B(x)^2), \quad (2)$$

$$A_m = \int \frac{S_0(\xi)\sigma_m^{SB}(\xi)}{D(\xi)} d\xi - k_1 \sum_i A_i \int \frac{\sigma_i^{SB}(\xi)\sigma_m^{SB}(\xi)}{D(\xi)} d\xi - k_2 \sum_i B_i \int \frac{\sigma_i^B(\xi)\sigma_m^{SB}(\xi)}{D(\xi)} d\xi, \quad (3)$$

$$B_m = \int \frac{S_0(\xi)\sigma_m^B(\xi)}{D(\xi)} d\xi - k_1 \sum_i A_i \int \frac{\sigma_i^{SB}(\xi)\sigma_m^B(\xi)}{D(\xi)} d\xi - k_2 \sum_i B_i \int \frac{\sigma_i^B(\xi)\sigma_m^B(\xi)}{D(\xi)} d\xi. \quad (4)$$

Here, $S_0(x)$ and $B_0(x)$ are the signal and background densities, respectively, in absence of systematic uncertainties. The constants k_1 and k_2 are discussed in detail in [1]. For the Banff challenge, $k_1 = 1$ and $k_2 = 1$ are used. The coefficients A_m and B_m are related to the systematic uncertainties and are obtained by solving the set of linear equations given above. The ξ integrals run over all valid marks. The systematic uncertainties from a source m are described by variations of the signal and background densities. The background density variation has the amplitude $\sigma_m^B(x)$. The density for signal-plus-background is varied with amplitude $\sigma_m^{SB}(x)$. The resulting background and signal-plus-background densities are given by

$$B(x) = B_0(x) + \sum_m f_m \sigma_m^B(x), \quad (5)$$

$$S(x) + B(x) = S_0(x) + B_0(x) + \sum_m f_m \sigma_m^{SB}(x), \quad (6)$$

where the f_m are nuisance parameters drawn from Gaussians with mean zero and width one. The Gaussians may be truncated in order to obtain positive densities $B(x)$ and $S(x) + B(x)$. The effect of truncating is not included in the definition of the weights $w(x)$.

In many cases the signal and background densities are not given in analytic form and only Monte Carlo experiments are available. The marks are then grouped into bins. For each bin there may be a bin-to-bin uncorrelated (statistical) error. These uncertainties enter as $\sigma_{unc}^B(x)$ and $\sigma_{unc}^S(x)$, for the background and signal shapes, respectively. For unbinned tests, $\sigma_{unc}^B = \sigma_{unc}^S(x) = 0$.

Finally, for a given experiment, the weights are summed to obtain the test statistics X:

$$X = \sum_i w(x_i). \quad (7)$$

The sum runs over all marks x_i observed in a single experiment.

For binned experiments, the marks are grouped into bins k with multiplicity N_k and bin weight w_k. The test statistics is then given by

$$X_{\rm binned} = \sum_k N_k w_k, \tag{8}$$

and the integrals given above simplify to sums over bins, for example

$$\int \frac{S_0(\xi)\sigma_m^{SB}(\xi)}{D(\xi)} d\xi \to \sum_k \frac{S_k^0 \sigma_{m,k}^{SB}}{D_k}. \tag{9}$$

2 Testing

For testing a single experiment, its p-value is calculated. The test statistics is calculated for the data, $X_{\rm data}$, and, repeatedly, for a sample of background experiments, calculated with Monte Carlo techniques. The p-value is given by

$$p = \frac{N(X_B \geq X_{\rm data})}{N_0}, \tag{10}$$

where N_0 is the total number of background experiments and $N(X_B \geq X_{\rm data})$ is the number of background experiments with $X_B \geq X_{\rm data}$. A positive decision is taken if $p < \alpha_l$, where α_l is the type 1 error.

3 Testing with unknown signal rate

In case of the Banff 2a challenge, the signal rates are not known. The signal $S_0(x)$ and the systematic errors take the form

$$\begin{aligned} S_0(x) &= r\tilde{S}_0(x), & (11)\\ \sigma_{unc}^S(x) &= r\tilde{\sigma}_{unc}^S(x), & (12)\\ \sigma_m^{SB}(x) &= r\tilde{\sigma}_m^S(x) + \sigma_m^B(x). & (13) \end{aligned}$$

where r is a nuisance parameter controlling the signal rate. For testing, the parameter $r = r_{\rm test}$ and a threshold $X_{\rm cut}$ are adjusted such that

$$\begin{aligned} X_{\rm cut} - \langle X_B \rangle &= \langle X_{SB} - X_B \rangle \,\text{and} & (14)\\ N(X_B > X_{\rm cut})) &= \alpha_l N_0. & (15)\\ & & (16) \end{aligned}$$

The average values of the test statistics in absence of a signal $\langle X_B \rangle$ or in presence of a signal $\langle X_{SB} \rangle$ depend on the choice of r through the definition of the event weights. They are calculated as

$$\langle X_B \rangle = \int B_0(\xi) w(\xi) d\xi, \tag{17}$$

$$\langle X_{SB} - X_B \rangle = \int S_0(\xi) w(\xi) d\xi. \tag{18}$$

It should be noted that this calculation neglects the effects of possibly truncated (asymmetric around S_0 or B_0) systematic uncertainties.

The test is performed by fixing the rate $r_{\rm test}$ and deciding on the basis of the p-value as described above. For a positive decision, the signal rate and a confidence interval are estimated by solving

$$X_{\rm data} - \langle X_B \rangle + \Delta = \langle X_{SB} - X_B \rangle. \tag{19}$$

for r, using certain choices of Δ. First, Δ is set to zero and an initial rate estimate r_0 is calculated. This rate estimate is biased, because the p-value requirement preferentially selects signal experiments with high $X_{\rm data}$. A bias-corrected signal parameter r is estimated from equation 19 with $\Delta = \Delta^{\rm bias}$ and $X^{r_0}_{\rm cut}$ defined as

$$\Delta^{\rm bias} = \frac{r}{r_0}(\langle X_{SB}(r_0)\rangle - \langle X_{SB}(r_0)\rangle_{X \geq X^{r_0}_{\rm cut}}), \tag{20}$$

$$N(X_B > X^{r_0}_{cut}) = \alpha_l N_0. \tag{21}$$

Here, the average X_{SB} is evaluated using signal-plus-background Monte Carlo events, once with and once without cutting events below $X^{r_0}_{\rm cut}$. Finally, a 68% confidence level $[r^-; r^+]$ is evaluated again using formula 19 with

$$\Delta^{\pm} = \frac{r^{\pm}}{r}(\Delta^{\rm bias} \pm {\rm RMS}(X_{SB})). \tag{22}$$

4 Testing with unknown parameter E

In case of the Banff 2a challenge, problem 1, in addition to the unknown signal rate, the parameter E is not known and shall be estimated. This is solved by scanning the parameter E in fine steps. Because the weight $w(x)$ suppresses contributions far from the signal peak, the "local" type 1 error rate α_l (for one scan point) needs to be adjusted, in order to match a given type 1 error for the full scan.

First, the rate $r_{\rm test}(E)$ is determined for each scan point E, as described above. The local type1 error α_l is adjusted using background Monte Carlo (MC) experiments: for each MC event a scan of the parameter E is performed, and the MC event is flagged if it is accepted as signal for any of the scan points. The global type 1 error is determined by counting the fraction of experiments which have been flagged. The local type 1 error is altered and the test is repeated until the desired global type 1 error is reached.

Finally, the data experiment is scanned, using the parameter α_l and the rate parameters $r_{\rm test}(E)$. In case of a positive decision, the properties of E are estimated: for each scan point E the significance

$$s(E) = \frac{X_{\rm data} - \langle X_B \rangle}{{\rm RMS}(X_B)} \tag{23}$$

is determined. The function $s(E)$ is approximated by a spline, and the maximum is determined. The 68% confidence level is estimated by determining the points where $s(E)$ changes by $-1/2$. Finally, after adjusting E, the signal rate and its confidence level are estimated as discussed above.

5 Results

Problem 1 is analysed once in 75 bins of the mark and once with an unbinned test. For the scan in E a total of 100 equidistant points are used. Problem 2 is analysed once in in 25 bins and once in 50 bins of the mark.

The fraction of signal events accepted by the algorithm is tested using 10000 signal-plus-background experiments. The results are summarised in table 1 for problem 1 and in table 2 for problem 2. The fraction of events where the estimated confidence interval contains the true value of the parameter is also given.

For problem 1, using 75 bins and $(D, E) = (1010, 0.1)$, a fraction of 37% is identified as signal. For $(D, E) = (137, 0.5)$ the fraction is 48% and for $(D, E) = (18, 0.9)$ it is 20%. The 68% confidence

Using 75 bins in the mark				
	total events	$p < \alpha_l$	E inside CL interval	D inside CL interval
$(D, E) = (1010, 0.1)$	10000	3727	3178 (85%)	2550 (68%)
$(D, E) = (137, 0.5)$	10000	4779	4015 (84%)	3492 (73%)
$(D, E) = (18, 0.9)$	10000	1971	1005 (51%)	1153 (58%)
Unbinned test				
	total events	$p < \alpha_l$	E inside CL interval	D inside CL interval
$(D, E) = (1010, 0.1)$	10000	3913	3316 (85%)	2373 (61%)
$(D, E) = (137, 0.5)$	10000	4712	3951 (84%)	2878 (61%)
$(D, E) = (18, 0.9)$	10000	1971	908 (46%)	1210 (61%)

Table 1: Results of the Banff 2a challenge, problem 1.

Using 25 bins in the mark			
	total events	$p < \alpha_l$	r inside 68% CL interval
rate $r = 25$	10000	1229	748 (61%)
rate $r = 50$	10000	4971	3866 (78%)
rate $r = 75$	10000	8540	5849 (68%)
rate $r = 100$	10000	9811	6466 (66%)
Using 50 bins in the mark			
	total events	$p < \alpha_l$	r inside 68% CL interval
rate $r = 25$	10000	1281	745 (58%)
rate $r = 50$	10000	4912	3823 (78%)
rate $r = 75$	10000	8522	5782 (68%)
rate $r = 100$	10000	9792	6434 (66%)

Table 2: Results of the Banff 2a challenge, problem 2.

levels in E are not defined very well. Obviously the simple algorithm of estimating the confidence interval from the significance function is not satisfactory. The confidence levels in D are surprisingly good, given the simplistic ansatz for the bias correction. The results obtained for problem 1 with an unbinned test are rather similar to the binned case.

For problem 2, using 25 bins and a rate of 75 events, 85% of the events are found as signal. Again, the algorithm to define the confidence level in the rate seems to work reasonably well. Using 50 instead of 25 bins does not change the results significantly.

6 Output files format

The Banff 2a challenge experiments have been analysed. For problem 1 (`p1binned_bias1_test.vec` and `p1unbinned_bias1_test.vec`), the lines of the output files have the format:

```
pval dec e e0 e1 d0 d1
```

where `pval` is the minimum p-value found while scanning the parameter E, `dec` is one if a positive decision is taken and zero otherwise. For a positive decision, `e` is the best guess of the parameter E, `e0` and `e1` define the estimated 68% CL interval for E. Finally, `d0` and `d1` define the estimated 68% CL interval for the parameter D. In case of a negative decision, the variables related to D and E are set to zero.

For problem 2 (`p2_25bins_bias1_test.vec` and `p2_50bins_bias1_test.vec`), each line of the output file is formatted as:

```
pval dec r r0 r1
```

Here, `pval` is the p-value, `dec` is one if a positive decision is taken and zero otherwise, `r` is the best guess of the signal rate, `r0` and `r1` define the 68% CL interval.

7 Software overview

The algorithms are implemented in C++, using the Root package. For binned tests, the user supplies the signal and background shapes, as well as systematic uncertainties in the form of Root histograms. The nuisances (systematic uncertainties) act on the normalisation of the histograms. For cases where the signal shape depends on nuisances, a method to generate the signal shape histogram for a given set of nuisances has to be implemented in addition.

For unbinned tests, the signal and background shapes as well as systematic uncertainties are provided as Root TF1 functions of the mark, the parameters taken from the nuisances. In addition to providing these functions, methods to generate signal and background events have to be implemented for the unbinned case.

The software is available here:
`www.desy.de/~sschmitt/Banff2aChallenge/Banff2aChallengeSschmitt.tgz`

References

[1] P. Bock, JHEP **0701** (2007) 080 [arXiv:hep-ex/0405072].